DATA SELECTION FOR BROADCAST NEWS CSR EVALUATIONS

William M. Fisher, Walter S. Liggett, Audrey Le, Jonathan G. Fiscus, and David S. Pallett

National Institute of Standards and Technology (NIST)
Room A216, Building 225 (Technology)
Gaithersburg, MD 20899

ABSTRACT

Composition of the 1997 Hub-4 broadcast news test set is discussed. The composition is based on concurrent selection of a statistically-equivalent test set for a future evaluation, adjustment of the set to match the training data, and other considerations. This paper discusses both the principles involved and the specific algorithms used.

1. INTRODUCTION

The use of speech from radio and television news broadcasts for the DARPA CSR (continuous speech recognition) evaluations suggests some new ways of thinking about these evaluations. As a test of systems intended to handle unanticipated speakers, large vocabularies, and other aspects of "found" speech, news broadcasts involve new challenges as well as most facets of speech recognition that were the object of previous CSR evaluations[1]. Thus, one might reasonably think of news broadcasts as a class of material encompassing enough to be the focus of a CSR research effort. A statistician might conceive of this class of material as a population of units from which probability samples can be drawn. With this framework, one can think about comparing systems through the use of independent samples from the population and about obtaining estimates of performance that apply to the entire population.

Typically, each year, every site is tested with a standardized common test set. Because this is appropriate for system comparisons, there is no reason for changing this aspect of the evaluation design. However, thinking about news broadcasts as a population suggests ways to answer new questions by changing other aspects of the design. In particular, a new question addressed by the design adopted this year is year-to-year comparison of system performance.

To think about obtaining a probability sample from a population, one has to think of the population as composed of units. Here, we think of news broadcasts as composed of topical units; in other words, stories. Moreover, one has to think about a list of all the units in the population or something equivalent.. Such a list is called a frame. A probability sample is a set of units chosen from the frame according to a random mechanism that accords every unit in the population a non-zero probability of inclusion. A simple random sample is a particular type of probability sample. A simple random sample is usually drawn unit by unit with each draw giving equal probability to the units in the population not already included in the sample. The data selected for the CSR evaluations discussed here consist of two simple random samples, although other types of probability samples might also have been appropriate.

As an illustration of the use of a probability sample, consider the choice of training data. If one wanted to train a CSR system for news broadcasts, one might request a simple random sample from the population of news broadcasts to use as training data. (This, of course, is not how the training data for the DARPA CSR evaluations were chosen, for good and practical reasons.) As an example, say that a network wanted to transcribe the last 50 years of its news broadcasts. One would not select a simple random sample of stories for this purpose because a list of all stories in the 50 years would not be available. One might select a simple random sample of news shows and then, from each, select a simple random sample of stories. This is called cluster sampling. It gives, not a simple random sample, but a probability sample. Such a sample would provide, after human transcription, training data that is representative in a specific sense.

Another application is choice of the evaluation data, although again this is not how it was done for the DARPA CSR evaluations. The purpose of the evaluation is comparison of the performance of CSR systems intended for news broadcasts. Thus, in somewhat ambiguous terms, one would like evaluation data that are representative of news broadcasts. If the evaluation data were selected arbitrarily, then the system best tuned to the population of news broadcasts might not perform most favorably. In an unambiguous way, a simple random sample is representative.

The application considered here is the selection of two sets of evaluation data as two independent random samples from arbitrarily selected material. For the purpose of the 1997 evaluation, the LDC set aside roughly 10 hours of news broadcasts. Three hours of this material were used for the 1997 evaluation. In selecting these 3 hours, we have made provision for comparison of the 1998 evaluation with the 1997 evaluation. We have done this by selecting another 1.5 hours of material for use in 1998.

The balance of the 1998 evaluation data will be made up of new material. Because the reuse of the same material two successive years is not advisable, we must select comparable but non-overlapping samples for the two years so that system improvements during the year can be tracked. We obtain comparable samples by randomly selecting units (stories) from the original 10 hours and randomly assigning them to the 1997 or 1998 evaluation data set.

It should be mentioned that whatever peculiarities the 10 hours of material have will carry over to the two random samples selected. In other words, if the 10 hours of material unfairly favor a particular system because the material is not representative of news broadcasts, then the two samples will also. One consequence of this is that the 1997 evaluation data will be particularly appropriate as training data for the 1998 evaluation. Perhaps evaluation participants can take advantage of this, but even if they do, the benefit will be limited because the 1998 evaluation will also contain new material.

2. STATISTICAL APPROACH

Use of independent random samples for the 1997 and 1998 evaluations allows us to determine whether the year-to-year performance difference can be attributed solely to the difference between the samples. Two independent random samples from the same population have a probability relationship that enables assessment of the effect of the sample difference on the performance difference.

One way to draw two random samples from a population is to draw one sufficiently large random sample and then subsample it. The 10 hours of material set aside for the 1997 evaluation cannot be conceived of as a true random sample from the population of news broadcasts. Nevertheless, in order to provide a framework for statistical inference, we pretend that after some editing, these 10 hours are a random sample from some population, although we know that this population is not the one we would have specified had we the choice unfettered by the practical considerations that governed the actual selection. We subsample this sample to obtain the 1997 evaluation data set and part of the 1998 set.

As an illustration, consider two evaluations of the same system performed with two non-overlapping test sets that are independent random samples. A statistical inference that is both simple and of interest is the test of the null hypothesis that the system did not change between evaluations. The data for this hypothesis test are the word error rates (and the word counts) for the units (stories) in the two samples. If the system did not change, the entire collection of word error rates (and word counts) would be a random sample from some probability distribution. The unit-to-unit variation in recognition difficulty is reflected in this distribution. Thus, under the null hypothesis, the distribution of the difference between the results of the two evaluations can be derived, and whether the difference between the two evaluations can reasonably be attributed to test set differences alone (with no system differences) can be determined.

The formulas needed to perform this hypothesis test can be obtained from Cochran (1977)[2]. These formulas involve an approximation based on the number of units in each sample being sufficiently large. One might answer the question of how many are sufficient by guessing tens of units (stories) in each sample or by pursuing a more detailed investigation. Let the number of units be n_1 for the first sample and n_2 for the second. For unit i of sample j, let the number of errors be y_{ij} and the number of words be x_{ij} so that the word error rate for the story is y_{ij}/x_{ij}. The word error rate for sample j is

$$\hat{R}_j = \sum_{i=1}^{n_j} y_{ij} \; / \; \sum_{i=1}^{n_j} x_{ij}.$$

The hypothesis test is based on the difference $\hat{R}_1 - \hat{R}_2$. Let

$$\overline{x} = \sum_{j=1}^{2} \sum_{i=1}^{n_j} x_{ij} / (n_1 + n_2)$$

The variance of the difference under the null hypothesis that the system did not change between evaluations can be estimated by

$$(\hat{R}_1 - \hat{R}_2) = \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \frac{1}{\overline{x}^2 (n_1 + n_2 - 1)} \sum_{j=1}^{2} \sum_{i=1}^{n_j} (y_{ij} - \hat{R}_j x_{ij}$$

Under the null hypothesis, the variance estimated by this formula accounts for the observed difference between the word error rates for the two evaluations. If the Type I error is taken to be 0.05, the critical point for the difference is

$$1.96 [v(\hat{R}_1 - \hat{R}_2)]^{1/2},$$

provided that $n_1 + n_2$ is large enough.

The size of this critical point can be estimated from the results of the 1996 evaluation. The number of units in the 1996 evaluation is 74 and the length of the test set is 2 hours. Assuming that the same ratio of number of units to duration holds, for the comparison of the 1997 evaluation (3 hours) with the 1998 evaluation (1.5 hours), the critical point will be approximately 5 percent. In other words, an absolute change in the word error rate of more than 5 percent due to system changes will be distinguishable from the year-to-year change in the evaluation test set.

3. SATISFYING OTHER CRITERIA

We could have listed the units (stories) in the 10 hours of material designated for the 1997 evaluation and selected two random samples from this list. We could have chosen units one after the other until samples of 3 hours and 1.5 hours were obtained. However, we did not, because we wanted to eliminate commercials and sports stories, to segregate a few very long stories, to more closely match the characteristics of the training data, and to minimize the number of newly-introduced inter-story transitions.

3.1 No Commercials and Sports Reports

From the list of units in the 10 hours, we simply deleted the commercials and sports reports. This is in accord with the test set specification.

3.2 Three Long Stories Treated Specially.

The list contains three stories with length greater than 15 minutes. These stories are all from C-SPAN *Washington Journal*. The problem with these three stories is that they contribute heavily to the word error rate because they are so long and they may be highly unusual, as can be inferred from their source and their length. Certain formulas used in this analysis are inaccurate for very skewed distributions; thus, if these stories were included in the list from which the test sets were selected, they might have distorted the statistical comparison of the two years, as outliers sometimes do. One way in which these stories are highly unusual is that they are rich in speech that can be classified as spontaneous broadcast speech (F1). For these reasons, we decided to truncate each of these stories to 15 minutes and to include the truncated stories in the test sets but not to include them simply randomly. We chose two to include in the 1997 set and one to include in the 1998 set. When we test statistical significance of the difference between 1997 and 1998, we will not include these stories. Otherwise, they will be included.

3.3 Pool Balanced Re Training Data

We further reduced the list to make the as-yet-to-be-drawn test sets a better match to the training data. In general terms, this balancing of the pool was done by a greedy algorithm which iteratively selected from the list of possible basic next moves the one which made the greatest difference in an objective function, until a stopping criteria was met.

In our particular case, a basic move was the de-selection of a unit (story). The objective function, which the algorithm tended to minimize, was a measure of the aggregate discrepancy between the resulting pool and the training set on features that we had some reason to believe affect the inherent difficulty of the speech and for which measurements were at hand, e.g., the focus conditions. Again after some experimentation, the actual function used was a weighted sum of differences in the percentage of time devoted to the focus conditions and the balance between radio and television. Table 1 below shows these weights; as can be seen, substantial weight was put on only the balance in F0, F1, and F2, with the major weight put on F1 balance.

Our target distribution of these features was not exactly that found in the training data. Our goal was to have the proportion of time in the weighted focus conditions and broadcast types be the same for the training data and for test sets with the 15-minute truncated stories restored. From this goal, we derived the proportions that we wanted the balanced test pool to have. Let the time proportion of F0 in the training data be p_{0T}, in the three 15-minute stories be p_{0L}, and in the target list be p_{0S}. Since the length of the three 15 minute stories is 45 minutes and the total length of both test sets is 270 minutes, we have as our target for reducing the list:

$$p_{0S} = \frac{6}{5} p_{0T} - \frac{1}{5} p_{0L}$$

A similar equation holds for the other features.

FEATURE	WEIGHT
% F0	0.27
% F1	0.46
% F2	0.21
% F3	0.01
% F4	0.01
% F5	0.01
% F6	0.01
% F7	0.01
% Radio	0.01
% TV	0.01

Table 1. Weights Used in Balancing.

The algorithm stopped when either the objection function had been reduced to 0.01 or less or when 50 iterations had occurred.

In general, we found that this process produces a reasonably close match between the training data and the test sets.

3.4 Sampling & Transition Smoothing.

The foregoing reductions in the list of units do not prevent us from regarding the list as a random sample from some population. Thus, we could have sub-sampled this list randomly to obtain simple random samples for the 1997 and 1998 evaluations. For statistical purposes, we will proceed as though we did this. In truth, we made an effort to make the transitions between stories more natural. Having selected a sample of stories from the list, we present them in the test set in the order in which they appeared in the original 10 hours. In this case, some pairs of stories are consecutive in the original 10 hours. The transitions between such pairs can be left as in the original material; in other words, in their natural state. Other pairs of stories require deletion of intervening material. We modified the signal between such stories so that the transition would not seem too unnatural. In addition, we made an effort to reduce the number of transitions of this second type by drawing not single stories but between one and five consecutive stories. The number of consecutive stories selected was chosen at random. We will not take this deviation from simple random

sampling into account in the statistical analysis because we believe that the effect will not be noticeable in the statistical results.

4. TEST SET CHARACTERISTICS

In this first exercise of automatic balancing, most weight was put on balance of the first three focus conditions. Figure 2 at the bottom of the page is a plot of the percentage of time devoted to the different focus conditions in each of several data sets, which is useful in understanding the main effects of the different processing steps. (Because these processing steps were done before final reconciliation of the reference annotations and transcriptions, the composition of the final test set varies slightly from what is shown here.) Several points are illustrated by this figure.

Note that in general, the pruning of the test pool by our balancing did make its distribution of test time across focus conditions more like the training set distribution.

Beside the bars for each focus condition is a vertical line with an "X" that shows the range and mean value of a set of 100 runs of our random sampling procedure. Note that this distribution is centered well at the value of the pruned test pool, as it should be.

And finally, notice that the bar representing our final random selection from the 100 runs is representative of the distribution, and not an outlier in any sense.

Figure 3 below similarly shows the changes in percentage of time devoted to the two broadcast types, radio and television. The weights on these features were nearly zero, so there was little or no control on their distribution in the balancing. The balancing does in fact move the distribution slightly farther away from the training set targets.

Due to schedule constraints, the balancing of the test data pool from which a random selection was made was based on preliminary annotations of the test data by one annotator. Subsequently a reconciliation process to correct the annotation was performed, in which three independent annotations were obtained and disagreements among them adjudicated. Thus the distribution of

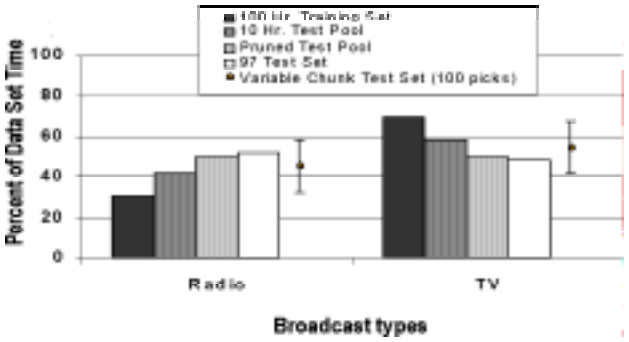

Figure 3. Broadcast Type Composition of Test Data.

the focus conditions in the final reconciled test data was somewhat different than determined by the random selection algorithm. Figure 4 below shows the before- and after-reconciliation percentage of time in the focus conditions, along with the repeated I-plots showing the distribution of trial picks made by the selection

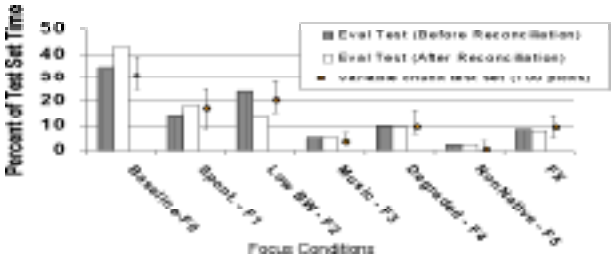

Figure 4. Effect of Reconciliation on focus Distribution .

program..

REFERENCES

1. Pallett, D.S., Fiscus, J.G., and Przybocki, M.A., "1996 Preliminary Broadcast News benchmark Tests,", in *Proc. Speech Recognition Workshop February 2-5, 1997*, Westfields International Conference Center, Chantilly, VA.

2. Cochran, W.G., *Sampling Techniques*, New York: John Wiley and Sons (1977).

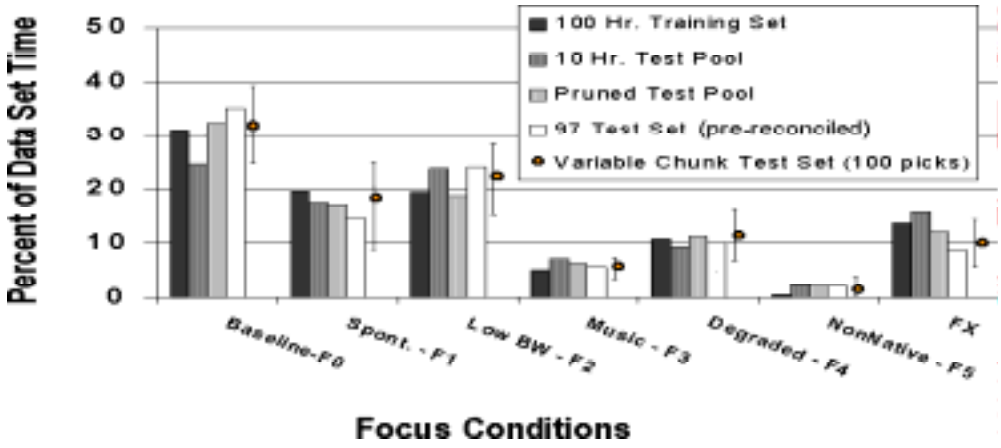

Figure 2. Focus Condition Composition of Test Data.